STAT 574 Linear and Nonlinear Mixed Models

Lecture 1: Review

Chencheng Cai

Washington State University

Linear Algebra

Vector Space (over real field)

A set V is a **vector space** if the followings hold for any $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w} \in V$ and $a, b \in \mathbb{R}$

- (closed under addition) $\boldsymbol{u} + \boldsymbol{v} \in V$.
- (closed under scalar multiplication) $a\boldsymbol{u} \in V$.
- (abelian group under addition)
  - (associativity) $(\boldsymbol{u} + \boldsymbol{v}) + \boldsymbol{w} = \boldsymbol{u} + (\boldsymbol{v} + \boldsymbol{w})$
  - (commutativity) $\boldsymbol{u} + \boldsymbol{v} = \boldsymbol{v} + \boldsymbol{u}$
  - (existence of identity) $\exists\ \mathbf{0} \in V, \boldsymbol{v} + \mathbf{0} = \boldsymbol{v}$ for all $\boldsymbol{v} \in V$.
  - (existence of inverse) For any $\boldsymbol{u} \in V$, there exists $-\boldsymbol{u} \in V$ such that $\boldsymbol{u} + (-\boldsymbol{u}) = \mathbf{0}$.
- (scalar multiplication)
  - $a(b\boldsymbol{u}) = (ab)\boldsymbol{u}$
  - $1\boldsymbol{u} = \boldsymbol{u}$
- (linear space)
  - $a(\boldsymbol{u} + \boldsymbol{v}) = a\boldsymbol{u} + a\boldsymbol{v}$
  - $(a + b)\boldsymbol{u} = a\boldsymbol{u} + b\boldsymbol{u}$

linear indpendence

- $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n \in V$ are **linearly independent** if the only solution to

$$a_1\boldsymbol{u}_1 + a_2\boldsymbol{u}_2 + \cdots + a_n\boldsymbol{u}_n = \boldsymbol{0}$$

  is $a_1 = a_2 = \cdots = a_n = 0$. Otherwise, they are **linearly dependent**.
- $\{\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_n\} \subseteq S$ is called the **maximal linearly-independent subset** of $S \subseteq V$ if for any $\boldsymbol{v} \in S$, $\{\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_n, \boldsymbol{v}\}$ are linearly dependent.
- The cardinality (size) of the maximal linearly-independent subset of $S \subseteq V$ is called the **rank** of S.

subspace and spanning

- $S \subseteq V$ is called a (linear) **subspace** of V if S inheritates the addition and the scalar multiplication from V and S itself is a vector space.
- The (linear) **span** of $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n\}$ is the smallest subspace of V that contains $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n\}$.

basis and dimension

- $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n\}$ is a **basis** of V if its elements are linearly independent and span the space V.
- The cardinality of any basis of V is the **dimension** of V.
- Let $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n\}$ be a basis of V. For any $\boldsymbol{v} \in V$, the decomposition

  $$\boldsymbol{v} = a_1\boldsymbol{u}_1 + a_2\boldsymbol{u}_2 + \cdots + a_n\boldsymbol{u}_n$$

  is unique, and the coefficients $a_1, \ldots, a_n$ are called the **coordinates** of $\boldsymbol{v}$ on the basis.
- Example: Euclidean space.

Inner Product Space

vector space + inner product = inner product space

- inner product $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{R}$.
  - $\langle \boldsymbol{u}, \boldsymbol{u} \rangle \geq 0$ and $\langle \boldsymbol{u}, \boldsymbol{u} \rangle = 0$ if and only if $\boldsymbol{u} = \mathbf{0}$
  - $\langle \boldsymbol{u}, \boldsymbol{v} \rangle = \langle \boldsymbol{v}, \boldsymbol{u} \rangle$
  - $\langle a\boldsymbol{u}, \boldsymbol{v} \rangle = a\langle \boldsymbol{u}, \boldsymbol{v} \rangle$
  - $\langle \boldsymbol{u}, \boldsymbol{v} + \boldsymbol{w} \rangle = \langle \boldsymbol{u}, \boldsymbol{v} \rangle + \langle \boldsymbol{u}, \boldsymbol{w} \rangle$
- inner product space is a normed space equipped with norm

$$\|\boldsymbol{u}\|_{\langle \cdot, \cdot \rangle} = \sqrt{\langle \boldsymbol{u}, \boldsymbol{u} \rangle}$$

Orthogonality

- $\boldsymbol{u} \neq \mathbf{0}$ and $\boldsymbol{v} \neq \mathbf{0}$ are othogonal if and only if $\langle \boldsymbol{u}, \boldsymbol{v} \rangle = 0$.
- A basis is **orthogonal** if its elements are pair-wise orthogonal.
- An orthogonal basis is **orthonormal** if any of the elements has norm 1.
- A mapping $P : V \to U \subset V$ is an **orthogonal projection** if and only if
  - $P\boldsymbol{u} = \boldsymbol{u}$ for any $\boldsymbol{u} \in U$.
  - $\langle P\boldsymbol{u}, \boldsymbol{u} - P\boldsymbol{u} \rangle = 0$ for any $\boldsymbol{u} \in V$.

Matrix

- Matrix is an array of real numbers:

$$\boldsymbol{A} = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m,1} & a_{m,2} & \cdots & a_{m,n} \end{bmatrix}$$

- Matrix is an aggregation of Euclidean vectors: $(\boldsymbol{u}_j \in \mathbb{R}^m)$

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{u}_1 & \boldsymbol{u}_2 & \ldots & \boldsymbol{u}_n \end{bmatrix}$$

- Matrix is a linear mapping:

$$\boldsymbol{A} : \mathbb{R}^m \to \mathbb{R}^n, (x_1, \ldots, x_m) \mapsto \left(\sum_{j=1}^{n} a_{1j} x_j, \ldots, \sum_{j=1}^{n} a_{nj} x_j \right)$$

We will skip..

- Basic operations of matrix.
- Special matrices (zero, identity, diagonal, etc..)
- Determinant.

Rank

If $\boldsymbol{A} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n] = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m]^T$, where $\boldsymbol{u}_j$'s are columns and $\boldsymbol{v}_i$'s are rows of $\boldsymbol{A}$, then

- $\text{span}(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n)$ is the column space or the manifold of $\boldsymbol{A}$, denoted by $\text{col}(\boldsymbol{A})$.
- rank:
$$\text{rank}(\boldsymbol{A}) := \text{rank}(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n) = \text{rank}(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m)$$
- rank is the dimension of the columns space.

$$\text{rank}(\boldsymbol{A}) = \dim(\text{col}(\boldsymbol{A})) = \dim(\text{col}(\boldsymbol{A}^T)) \leq m \wedge n$$

Trace

- Trace of a squared matrix is the sum of the elements on the diagnoal.

$$\mathrm{tr}(\boldsymbol{A}) = \sum_{i=1}^{n} A_{ii}$$

- Use trace to present sum of pairwise products of two matrices. Let $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{m \times n}$. Then we have

$$\mathrm{tr}(\boldsymbol{A}^T \boldsymbol{B}) = \mathrm{tr}(\boldsymbol{B}^T \boldsymbol{A}) = \sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij} B_{ij}$$

Moore-Penrose Inverse

- For $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, a psudo-inverse $\boldsymbol{A}^{+} \in \mathbb{R}^{n \times n}$ satisfies
  - $\boldsymbol{A}\boldsymbol{A}^{+}\boldsymbol{A} = \boldsymbol{A}$
  - $\boldsymbol{A}^{+}\boldsymbol{A}\boldsymbol{A}^{+} = \boldsymbol{A}^{+}$
  - Both $\boldsymbol{A}\boldsymbol{A}^{+}$ and $\boldsymbol{A}^{+}\boldsymbol{A}$ are symmetric.
- For $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ $(m > n)$, if $\mathrm{rank}(\boldsymbol{A}) = n$, then

$$\boldsymbol{A}^{+} = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T.$$

Woodbury Identity

- If $\boldsymbol{A}$ and $\boldsymbol{C}$ are invertible, and assuming all matrices are conformal, we have

$$(\boldsymbol{A}+\boldsymbol{U}\boldsymbol{C}\boldsymbol{V})^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{U}(\boldsymbol{C}^{-1}+\boldsymbol{V}\boldsymbol{A}^{-1}\boldsymbol{U})^{-1}\boldsymbol{V}\boldsymbol{A}^{-1}$$

- Special case: $\boldsymbol{A}=\boldsymbol{I}$, $\boldsymbol{C}=[1]$, $\boldsymbol{U}=\boldsymbol{V}^T=\boldsymbol{u}$.

$$(\boldsymbol{I}+\boldsymbol{u}\boldsymbol{u}^T)^{-1} = \boldsymbol{I} - \frac{\boldsymbol{u}\boldsymbol{u}^T}{1+\|\boldsymbol{u}\|^2}$$

- Special case: $\boldsymbol{U}=\boldsymbol{C}=\boldsymbol{I}$.

$$(\boldsymbol{A}+\boldsymbol{C})^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}(\boldsymbol{A}^{-1}+\boldsymbol{C}^{-1})^{-1}\boldsymbol{A}^{-1}$$

Eigenvalues and eigenvectors for symmetric matrices

Let $\boldsymbol{A}$ be an $n \times n$ symmetric matrix

- If $\boldsymbol{A}\boldsymbol{u} = \lambda \boldsymbol{u}$, then λ is called an **eigenvalue** of $\boldsymbol{A}$, and $\boldsymbol{u}$ is the **eigenvector**.
- $\boldsymbol{A}$ has n eigenvalues and eigenvectors (including zeros and duplicated eigenvalues).
- Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ be the eigenvalues in descending order, and $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ be the corresponding eigenvectors.
- If $\lambda_n > 0$, then $\boldsymbol{A}$ is positive-definite that $\boldsymbol{w}^T \boldsymbol{A} \boldsymbol{w} > 0$ for all $\boldsymbol{w} \in \mathbb{R}^n$ and $\boldsymbol{w} \neq \mathbf{0}$. If $\lambda \geq 0$, $\boldsymbol{A}$ is positive semi-definite.
- $\boldsymbol{A}$ is singular if and only if $\lambda_n = 0$.
- Rank of $\boldsymbol{A}$ equals the number of non-zero eigenvalues.

Eigenvalues and eigenvectors for symmetric matrices

- $\boldsymbol{u}_1$ is the optimum to the optimization:

$$\max_{\|\boldsymbol{w}\|=1} \; \boldsymbol{w}^T \boldsymbol{A} \boldsymbol{w}$$

- $\boldsymbol{u}_i$ $(i > 1)$ is the optimum to the optimization:

$$\max_{\|\boldsymbol{w}\|=1, \boldsymbol{w}^T \boldsymbol{u}_j = 0 \text{ for } 1 \leq j < i} \; \boldsymbol{w}^T \boldsymbol{A} \boldsymbol{w}$$

Eigenvalues Decomposition

- We can write
$$\boldsymbol{A} = \sum_{i=1}^{n} \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^T$$
- Or
$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^T,$$
where $\boldsymbol{D} = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ and $\boldsymbol{U} = [\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_n]$ is orthonormal.

Singular values and singular vectors for squared matrices

Let $\boldsymbol{A}$ be an $m \times n$ matrix with $m > n$.

- If $\boldsymbol{A}\boldsymbol{u} = s\boldsymbol{v}$ and $\boldsymbol{A}^T\boldsymbol{v} = s\boldsymbol{u}$, then s is a singular value of $\boldsymbol{A}$, and $\boldsymbol{u}$ and $\boldsymbol{v}$ are the right and left singular vectors.
- s^2 is an eigenvalue of $\boldsymbol{A}^T\boldsymbol{A}$ and $\boldsymbol{u}$ is the eigenvector.
- s^2 is an eigenvalue of $\boldsymbol{A}\boldsymbol{A}^T$ and $\boldsymbol{v}$ is the eigenvector.
- $\boldsymbol{A}$ has at most n non-zero singular values.
- Let the singular values be $s_1 \geq s_1 \geq \cdots \geq s_n$, and the singular vectors be $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$ for $i = 1, \ldots, n$.

Singular values and singular vectors for squared matrices

- $\boldsymbol{u}_1$ and $\boldsymbol{v}_1$ are the optimum to the optimization:

$$\max_{\|\boldsymbol{w}\|=1, \|\boldsymbol{z}\|=1} \boldsymbol{w}^T \boldsymbol{A} \boldsymbol{z}$$

- $\boldsymbol{u}_1$ is the optimum to the optimization:

$$\max_{\|\boldsymbol{w}\|=1} \boldsymbol{w}^T \boldsymbol{A}^T \boldsymbol{A} \boldsymbol{w}$$

- $\boldsymbol{v}_1$ is the optimum to the optimization:

$$\max_{\|\boldsymbol{w}\|=1} \boldsymbol{w}^T \boldsymbol{A} \boldsymbol{A}^T \boldsymbol{w}$$

Singular Value Decomposition

- We can write

$$\boldsymbol{A} = \sum_{i=1}^{n} s_i \boldsymbol{v}_i \boldsymbol{u}_i^T$$

- Or

$$\boldsymbol{A} = \boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^T,$$

where $\boldsymbol{D} = \mathrm{diag}(s_1, \ldots, s_n)$, $\boldsymbol{V} = [\boldsymbol{v}_1, \cdots, \boldsymbol{v}_n]$ and $\boldsymbol{U} = [\boldsymbol{u}_1, \cdots, \boldsymbol{u}_n]$. Both $\boldsymbol{U}$ and $\boldsymbol{V}$ are orthonormal.

Other Decompositions

- **Cholesky Decompositin.**
  If $\boldsymbol{A}$ is symmetric positive definite, then
  $$\boldsymbol{A} = \boldsymbol{L}\boldsymbol{L}^T$$
  for some lower triangular matrix $\boldsymbol{L}$.
- **LU Decomposition.**
  If $\boldsymbol{A}$ is a square matrix, then
  $$\boldsymbol{A} = \boldsymbol{L}\boldsymbol{U}^T$$
  for some lower triangular matrix $\boldsymbol{L}$ and some upper triangular matrix $\boldsymbol{U}$.
- **QR Decomposition.**
  If $\boldsymbol{A}$ is $m \times n$, then
  $$\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{R}$$
  for some orthogonal $m \times m$ matrix $\boldsymbol{Q}$ and some upper triangular $m \times n$ matrix $\boldsymbol{R}$.

Matrix Calculus

Basic definitions

- matrix calculus = multivariate calculus + assembling
- univariate scalar function: $f' = df/dx$
- multivariate scalar function:

$$\nabla f = \partial f/\partial \boldsymbol{x} = (\partial f/\partial x_1, \partial f/\partial x_2, \partial f/\partial x_3, \ldots, \partial f/\partial x_n)$$

- univariate vector function:

$$\boldsymbol{f}' = d\boldsymbol{f}/dx = (df_1/dx, df_2/dx, \ldots, df_k/dx)^T$$

- multivariate vector function:

$$\nabla \boldsymbol{f} = \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial f_k}{\partial x_1} & \frac{\partial f_k}{\partial x_2} & \cdots & \frac{\partial f_k}{\partial x_n} \end{bmatrix}$$

Basic definitions

- function is matrix-valued:

$$\frac{d\boldsymbol{M}}{dx} = \begin{bmatrix} \frac{dM_{11}}{dx} & \frac{dM_{12}}{dx} & \cdots & \frac{dM_{1n}}{dx} \\ \frac{dM_{21}}{dx} & \frac{dM_{22}}{dx} & \cdots & \frac{dM_{2n}}{dx} \\ \vdots & \vdots & & \vdots \\ \frac{dM_{m1}}{dx} & \frac{dM_{m2}}{dx} & \cdots & \frac{dM_{mn}}{dx} \end{bmatrix}$$

- function of matrices:

$$\frac{\partial f}{\partial \boldsymbol{X}} = \begin{bmatrix} \frac{\partial f}{\partial X_{11}} & \frac{\partial f}{\partial X_{12}} & \cdots & \frac{\partial f}{\partial X_{1n}} \\ \frac{\partial f}{\partial X_{21}} & \frac{\partial f}{\partial X_{22}} & \cdots & \frac{\partial f}{\partial X_{2n}} \\ \vdots & \vdots & & \vdots \\ \frac{\partial f}{\partial X_{m1}} & \frac{\partial f}{\partial X_{m2}} & \cdots & \frac{\partial f}{\partial X_{mn}} \end{bmatrix}$$

Differentiation

- univariate scalar function: $df = f' dx$
- multivariate scalar function:
$$df = \nabla f \, d\boldsymbol{x}$$
- univariate vector function:
$$d\boldsymbol{f} = \boldsymbol{f}' dx$$
- multivariate vector function:
$$d\boldsymbol{f} = \nabla \boldsymbol{f} \, d\boldsymbol{x}$$
- matrix-valued function:
$$d\boldsymbol{M} = \frac{d\boldsymbol{M}}{dx} dx$$
- function of matrices:
$$df = \operatorname{tr}\left[\left(\frac{\partial f}{\partial \boldsymbol{X}} \right)^T d\boldsymbol{X} \right]$$

Differentiation — expending to more components

- univariate scalar function: $df = f'_x dx + f'_y dy$
- multivariate scalar function:

$$df = \nabla_x f \, d\boldsymbol{x} + \nabla_y f \, d\boldsymbol{y}$$

- univariate vector function:

$$d\boldsymbol{f} = \boldsymbol{f}'_x dx + \boldsymbol{f}'_y dy$$

- multivariate vector function:

$$d\boldsymbol{f} = \nabla_x \boldsymbol{f} \, d\boldsymbol{x} + \nabla_y \boldsymbol{f} \, d\boldsymbol{y}$$

- matrix-valued function:

$$d\boldsymbol{M} = \frac{\partial \boldsymbol{M}}{\partial x} dx + \frac{\partial \boldsymbol{M}}{\partial y} dy$$

- function of matrices:

$$df = \operatorname{tr}\left[\left(\frac{\partial f}{\partial \boldsymbol{X}}\right)^T d\boldsymbol{X}\right] + \operatorname{tr}\left[\left(\frac{\partial f}{\partial \boldsymbol{Y}}\right)^T d\boldsymbol{Y}\right]$$

Chain Rules

Iteratively replace differentiations.

- Differentiation for $f(\boldsymbol{X}(t), \boldsymbol{Y}(t))$:

$$
\begin{aligned}
df &= \mathrm{tr}\left[\left(\frac{\partial f}{\partial \boldsymbol{X}}\right)^T d\boldsymbol{X}\right] + \mathrm{tr}\left[\left(\frac{\partial f}{\partial \boldsymbol{Y}}\right)^T d\boldsymbol{Y}\right] \\
&= \left\{\mathrm{tr}\left[\left(\frac{\partial f}{\partial \boldsymbol{X}}\right)^T \frac{d\boldsymbol{X}}{dt}\right] + \mathrm{tr}\left[\left(\frac{\partial f}{\partial \boldsymbol{Y}}\right)^T \frac{d\boldsymbol{Y}}{dt}\right]\right\} dt
\end{aligned}
$$

- Differentiation for $f(g(\boldsymbol{x}, z))$:

$$
df = f' dg = f'(\nabla_x g d\boldsymbol{x} + g'_z dz) = f' \nabla_x g d\boldsymbol{x} + f' g'_z dz
$$

Common Results

- Let $y = \boldsymbol{u}^T \boldsymbol{x}$.
$$\nabla y = \boldsymbol{u}^T$$
- Let $y = \boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x}$.
$$\nabla y = \boldsymbol{x}^T \boldsymbol{A} + \boldsymbol{x}^T \boldsymbol{A}^T$$
- Let $y = \boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x}$ with symmetric $\boldsymbol{A}$.
$$\nabla y = 2\boldsymbol{x}^T \boldsymbol{A}$$
- Let $y = \|\boldsymbol{x}\|^2$.
$$\nabla y = 2\boldsymbol{x}^T$$
- Let $y = \|\boldsymbol{x}\|$.
$$\nabla y = \frac{\boldsymbol{x}^T}{\|\boldsymbol{x}\|}$$
- Let $\boldsymbol{y} = \boldsymbol{A} \boldsymbol{x}$.
$$\nabla \boldsymbol{y} = \boldsymbol{A}$$

Common Results

- Let $y = \mathrm{tr}(\boldsymbol{A}^T \boldsymbol{X})$.

$$\frac{\partial y}{\partial \boldsymbol{X}} = \boldsymbol{A}$$

- Let $y = \mathrm{tr}(\boldsymbol{X})$.

$$\frac{\partial y}{\partial \boldsymbol{X}} = \boldsymbol{I}$$

- Let $y = \boldsymbol{u}^T \boldsymbol{X} \boldsymbol{v}$.

$$\frac{\partial y}{\partial \boldsymbol{X}} = \boldsymbol{u}\boldsymbol{v}^T$$

- Let $y = |\boldsymbol{X}|$.

$$\frac{\partial y}{\partial \boldsymbol{X}} = |\boldsymbol{X}|\boldsymbol{X}^{-1}$$

Multivariate matrix differentiation

- We know that

$$d(\boldsymbol{X}\boldsymbol{Y}) = (d\boldsymbol{X})\boldsymbol{Y} + \boldsymbol{X}d\boldsymbol{Y}$$

- Then

$$0 = (d\boldsymbol{X})\boldsymbol{X}^{-1} + \boldsymbol{X}d(\boldsymbol{X}^{-1})$$

- Therefore

$$d(\boldsymbol{X}^{-1}) = -\boldsymbol{X}^{-1}(d\boldsymbol{X})\boldsymbol{X}^{-1}$$

Example:
Let $y = \boldsymbol{u}^T(\boldsymbol{I} + x\boldsymbol{D})^{-1}\boldsymbol{v}$.

$$\begin{aligned} dy &= \boldsymbol{u}^T d(\boldsymbol{I} + x\boldsymbol{D})^{-1}\boldsymbol{v} \\ &= -\boldsymbol{u}^T(\boldsymbol{I} + x\boldsymbol{D})^{-1} d(\boldsymbol{I} + x\boldsymbol{D})(\boldsymbol{I} + x\boldsymbol{D})^{-1}\boldsymbol{v} \\ &= -\boldsymbol{u}^T(\boldsymbol{I} + x\boldsymbol{D})^{-1}\boldsymbol{D}(\boldsymbol{I} + x\boldsymbol{D})^{-1}\boldsymbol{v}dx \end{aligned}$$

Linear Regression

Linear Regression Model

- Coordinate-wise

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i \quad \text{for } i = 1, \ldots, n$$

- Vectorize independent variables

$$y_i = \boldsymbol{\beta}^T \boldsymbol{x}_i + \epsilon_i \quad \text{for } i = 1, \ldots, n$$

- Vectorize observations

$$\boldsymbol{y} = \beta_0 \mathbf{1} + \beta_1 \boldsymbol{x}^{(1)} + \beta_2 \boldsymbol{x}^{(2)} + \cdots + \beta_p \boldsymbol{x}^{(p)} + \boldsymbol{\epsilon}$$

- Matrix form

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

Notation

- x_{ij}: value of j-th independent variable of unit i.
- $\boldsymbol{x}_i := (1, x_{i1}, x_{i2}, \ldots, x_{ip})^T$: vector of indepednent variables of unit i.
- $\boldsymbol{x}^{(j)} := (x_{1j}, x_{2j}, \ldots, x_{nj})^T$: vector of j-th independent variable from all units.
- $\boldsymbol{X} := [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n]^T = [\mathbf{1}, \boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(p)}]$: design matrix.
- $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$: coefficient vector.
- $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \epsilon_n)^T$: noise/error vector.

Some useful identities:

- $\boldsymbol{X}^T\boldsymbol{X} = \sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^T$
- $[\boldsymbol{X}^T\boldsymbol{X}]_{jk} = \left[\boldsymbol{x}^{(j-1)}\right]^T \boldsymbol{x}^{(k-1)}$ by letting $\boldsymbol{X}^{(0)} = \mathbf{1}$.

Assumptions (LINE)

- **L**inear relationship between the mean response and the independent variables.
  diagnostics: scatter plot, partial regression plot.
- **I**ndependent observations. The errors ϵ_i's are independent.
- **N**ormally distributed. The errors ϵ_i's are normally distributed.
  diagnostics: QQ plot for residuals.
- **E**qual variances. The errors ϵ_i's have equal variances.
  diagnostics: residual plot.

In summary:

$$\boldsymbol{y} \sim \mathcal{N}_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n)$$

- Observed: $\boldsymbol{X}$, $\boldsymbol{y}$
- Unknown: $\boldsymbol{\beta}$, σ^2

Least Squares Estimation (LSE)

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2$$

- objective function: residual sum-of-squares.
- solution: $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$.
- requirement: $\boldsymbol{X}^T\boldsymbol{X}$ invertible.
- fitted value: $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} =: \boldsymbol{H}\boldsymbol{y}$ where $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ is called **hat matrix**.
- residual: $\hat{\boldsymbol{\epsilon}} = \boldsymbol{y} - \hat{\boldsymbol{y}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}$
- unbiased estimator for variance: $\hat{s}^2 = \|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2/(n - (p+1))$

Maximum Likelihood Estimation (MLE)

$$\max_{\boldsymbol{\beta},\sigma^2} \frac{1}{(2\pi)^{n/2}\sqrt{|\sigma^2 \boldsymbol{I}|}} \exp\left\{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T(\sigma^2\boldsymbol{I})^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\right\}$$

- solution:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$
$$\hat{\sigma}^2 = \frac{1}{n}\|\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2$$

- requirement: $\boldsymbol{X}^T\boldsymbol{X}$ invertible.

Distributions

$$\boldsymbol{y} \sim \mathcal{N}_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I})$$

- $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})$
- $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{H})$
- $\hat{\boldsymbol{\epsilon}} = \boldsymbol{y} - \hat{\boldsymbol{y}} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2(\boldsymbol{I} - \boldsymbol{H}))$
- $\|\hat{\boldsymbol{\epsilon}}\|^2 \sim \sigma^2\chi^2_k$ where $k = \mathrm{rank}(\boldsymbol{I} - \boldsymbol{H}) = n - p - 1$.
- Fact: if $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$ and $\boldsymbol{\Sigma}^2 = \boldsymbol{\Sigma}$, then $\|\boldsymbol{x}\|^2 \sim \chi^2_k$ where $k = \mathrm{rank}(\boldsymbol{\Sigma})$.

Gauss-Markov Theorem

- Under the conditions of linear regression model, $\hat{\boldsymbol{\beta}}$ is the best linear unbiased estimator (BLUE) for $\boldsymbol{\beta}$.
- That is if $\tilde{\boldsymbol{\beta}} = \boldsymbol{w}^T \boldsymbol{y}$ for some $\boldsymbol{w}$ and $\mathbb{E}[\tilde{\boldsymbol{\beta}}] = \boldsymbol{\beta}$, then

$$\mathrm{Var}(\tilde{\boldsymbol{\beta}}) \succeq \sigma^2 (\boldsymbol{X}^T \boldsymbol{X})^{-1}.$$

Multicollinearity

- Multicollinearity: near-perfect linear dependence among the predictors.
- Quantification: variance-inflation-factor (VIF).
- The issue:
  - $\boldsymbol{X}^T\boldsymbol{X}$ is close to be singular.
  - large variance for $\hat{\boldsymbol{\beta}}$.
- Solution:
  - Variable Selection: best subset, stepwise selection.
  - Penalized Linear Regression: ridge, LASSO.